A COMPLEX WAVELET BASED FUNDAMENTAL FREQUENCY ESTIMATOR IN SINGLE-CHANNEL POLYPHONIC SIGNALS

Jesús Ponce de León, *
mailto:jponce@unizar.es

Fernando Beltrán
mailto:beltran@unizar.es

José R. Beltrán,†
Dept. of Electronic Engineering and Communications
University of Zaragoza
Zaragoza, Spain
mailto:jrbelbla@unizar.es

ABSTRACT

In this work, a new estimator of the fundamental frequencies (F_0) present in a polyphonic single-channel signal is developed. The signal is modeled in terms of a set of discrete partials obtained by the Complex Continuous Wavelet Transform (CCWT). The fundamental frequency estimation is based on the energy distribution of the detected partials of the input signal followed by an spectral smoothness technique. The proposed algorithm is designed to work with suppressed fundamentals, inharmonic partials and harmonic related sounds. The detailed technique has been tested over a set of input signals including polyphony 2 to 6, with high precision results that show the strength of the algorithm. The obtained results are very promising in order to include the developed algorithm as the basis of Blind Sound Source Separation or automatic score transcription techniques.

1. INTRODUCTION

The perceptual property of pitch is one of the auditory attributes of musical tones, along with duration, loudness, and timbre. Since pitch may be quantified as a frequency, the automatic pitch detection has turned into the estimation of the fundamental frequency, F_0. The single-F_0 estimation is needed for example in speech recognition and music information retrieval. Existing algorithms for pitch estimation include, among others, the Average Magnitude Difference Function [1], Harmonic Product Spectrum [2], Parallel Processing Pitch Detector [3], Square Difference Function [4], Cepstral Pitch Determination [5], Subharmonic to harmonic ratio [6] and Super Resolution Pitch Detector [7].

Different techniques have been proposed for multiple F_0 estimation, associated with complex applications like automatic music transcription or blind audio source separation. In these cases, a multiple F_0 estimator is needed. The difficulty of multiple-F_0 estimation lies in the fact that sound sources often overlap in time as well as in frequency due to the specific nature of the different musical instruments or the characteristic of the environment (reverberation). Without any harmonic assumption, a sound can be mathematically defined [8] as:

$$x(t_n) = \sum_{i=1}^{M} A_i(t_n) cos[\theta_i(t_n)] + e(t_n) \quad (1)$$

where M is the total number of oscillators, t_n is the (discrete) time index, $A_i(t_n)$ and $\theta_i(t_n)$ are the (instantaneous) amplitude and phase of the i^{th} oscillator and $e(t_n)$ is the noise component.

On the other hand, most of the existing multipitch detection methods are based on harmonicity. In these methods, although assuming that some musical instruments (like piano) can have some degree of inharmonicity, a note played by a musical instrument is usually considered a harmonic sound source. The mixture of several instruments playing different notes can be modeled [9] as:

$$x(t_n) = \sum_{i=1}^{N} h_i(t_n) + e'(t_n) \quad (2)$$

where N is the number of sources (polyphony), $h_i(t_n)$ is the quasi-periodic part of the i^{th} source (that is, sources are supposed harmonic) and $e'(t_n)$ the non-harmonic or noisy part of the signal. Two important (and difficult) problems to deal with are the modeling of the harmonic part of the source, $h_i(t_n)$, and the decomposition of the mixed signal into an unknown number N of sources.

Some existing techniques use the sum of amplitudes to obtain a weight function capable of selecting the different F_0 present in the signal [10]. Other methods propose the use of several combined criteria in order to select the different F_0 candidates [9]. The spectral smoothing technique [11], [12] is proposed as an efficient mechanism in estimating the spectral envelopes of the detected sounds.

Most of the proposed methods use the Short Time Fourier Transform (STFT) to analyze the input signal in order to find the sinusoidal components and hence the harmonic part of the sound. In this work, a new approach will be used. Based on the Complex Continuous Wavelet Transform (CCWT), we have developed an analysis/synthesis algorithm, namely the Complex Wavelet Additive Synthesis (CWAS) algorithm [13], [14]. As will be shown is Section 2, the model of the audio signal proposed by the CWAS algorithm can be written as:

$$x(t_n) = \sum_{i=1}^{n} A_i(t_n) \cos[\Phi_i(t_n)] \quad (3)$$

* Corresponding author

† This work has been supported by the Spanish government project TEC2009-14414-C03-01 (*Analysis, Classification and Separation of Sound Sources, $AnClaS^3$ v2.0*).

where $A_i(t_n)$ and $\Phi_i(t_n)$ are the instantaneous amplitude and the instantaneous phase of each one of the n detected partials. Observe the similarities between this expression and Equation (2). In our model, we only have to extract the harmonic partials of the signal, being the sum of non-harmonic partials considered as noise.

From this model, a set of criteria can be used to find F_0 as the best candidate. The spectral smoothness method is used to subtract at least a part of the harmonic envelope of the detected source, iterating the process in order to estimate the next F_0 candidate. A previous version of this estimation algorithm was presented in [15] and used to separate the different mixed sources [14].

This paper is divided as follows: Section 2 provides a brief introduction to the CCWT and the CWAS algorithm, including the interpretation of their results and the additive synthesis process. The proposed F_0 estimator and its main blocks will be presented in Section 3. The first accuracy results from polyphony 2 to 6 are shown in Section 4. Finally, the main conclusions and current and future lines of work are presented in Section 5. This work has been developed in *MATLAB* ®.

2. COMPLEX BANDPASS FILTERING

The analysis of the audio signal is carried through a discrete version of the Complex Continuous Wavelet Transform (CCWT), the Complex Wavelet Additive Synthesis (CWAS) algorithm [13], [14]. The most common mathematical definition of the CCWT for a certain input signal $x(t)$ can be written as [16]:

$$W_x(a,b) = \int_{-\infty}^{+\infty} x(t)\Psi^*_{a,b}(t)dt \tag{4}$$

where * is the complex conjugate and $\Psi_{a,b}(t)$ is the *mother* or *atom wavelet*, temporally shifted by a factor b and frequency scaled by a factor a:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right) \tag{5}$$

Therefore, the wavelet transform of a signal is equivalent to a band pass filtering of the signal. We have developed an analysis algorithm based on the Morlet wavelet:

$$\Psi(t) \approx \; Ce^{-\frac{t^2}{2\sigma^2}}e^{j\omega_0 t} \tag{6}$$

In the time domain, the Morlet wavelet is a complex exponential modulated by a Gaussian of width $2\sqrt{2}\sigma$, centered on the frequency ω_0/a. Its Fourier transform is:

$$\hat{\psi}_a(\omega) = C'e^{-\sigma^2\frac{(a\omega-\omega_0)^2}{2}} \tag{7}$$

C and C' are the normalization constants of the mother wavelet in time and frequency domain, respectively [17]. Hence, the CCWT is equivalent to filtering the signal through a bandpass filter bank whose frequency response is given by Equation (7). The exact shapes of Equations (6) and (7) are shown in Figure 1.

In order to quantize the filter bank structure, we must use a discrete scale parameter k rather than a continuous a. As proposed by [16], [18], we used a dyadic set of scale factors k_n. This frequency division provides a logarithmic-resolution frequency axis. A new parameter can be introduced, the number of divisions per octave, D, which controls the resolution of the analysis. The set of discrete scales are then obtained by:

$$k_n = k_{min}2^{\frac{n}{D}}, n = 1,\ldots,N\cdot D \tag{8}$$

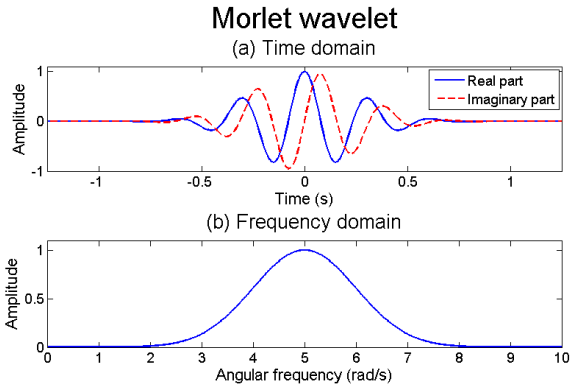

Figure 1: *The Morlet wavelet. (a) Time domain. In blue, real part. In red (dashed), imaginary part. (b) Frequency domain.*

If $D = 1$, the spectrum is divided into N octaves. The minimum scale k_{min} is related to the maximum frequency of the analysis, f_{max}, and f_{max} is related to the sampling rate f_s by the Nyquist criterion, $f_{max} = f_s/2$.

Hence, in Equation (7) the continuous scale factor a can be replaced for the discrete k_n, obtaining the quantized version of Morlet wavelet in the *scale* axis ($k = \omega/\omega_0$), which can be expressed as:

$$\hat{\psi}_{k_n}(k) = C_k e^{-\frac{\sigma^2\omega_0^2}{2}\left(\frac{k}{k_n}-1\right)^2} \tag{9}$$

Due to the complex nature of the transform, the wavelet coefficients can be written as:

$$W_x(k_n,b) = \begin{cases} IFFT[2\hat{x}(k)\hat{\psi}_{k_n}(k)] & \text{if } k>0 \\ IFFT[\hat{x}(0)\hat{\psi}_{k_n}(0)] & \text{if } k=0 \\ 0 & \text{if } k<0 \end{cases} \tag{10}$$

The result of filtering a signal through ψ_{k_n} is a matrix of complex numbers (the wavelet coefficients), $W_x(k_n,b)$, where k_n are the set of discrete scales of analysis and b is the discrete temporal variable (from now t). These coefficients can be studied in modulus and phase, being:

$$\|W_x(k_n,t)\| = \sqrt{\Re[W_x(k_n,t)]^2 + \Im[W_x(k_n,t)]^2} \tag{11}$$

$$\Phi_x(k_n,t) = arg\left(\Re[W_x(k_n,t)] + j\Im[W_x(k_n,t)]\right) \tag{12}$$

It is possible to obtain the instantaneous amplitude of the signal $A(t)$ from Equation (11) and its instantaneous phase $\phi(t)$ from Equation (12). The instantaneous frequency of the signal evaluated in the k_j scale can be obtained from the temporal derivative of Equation (12), [19]:

$$f_{ins}(k_j,t) = \frac{1}{2\pi}\frac{d[\Phi_x(k_j,t)]}{dt} \tag{13}$$

One example of the modulus of $W_x(k_n,t)$, also called the *wavelet spectrogram*, can be seen in Figure 2 (down), for a guitar playing a $E2$ note. The bright zones are related with the detected partials. The waveform of the input signal can be seen in the upper part of the figure.

The sum over the time axis of the module of the wavelet coefficients represents the *scalogram* of the signal. The scalogram presents a certain number of peaks, each one related to a detected partial of the signal. In the model of the audio signal proposed in

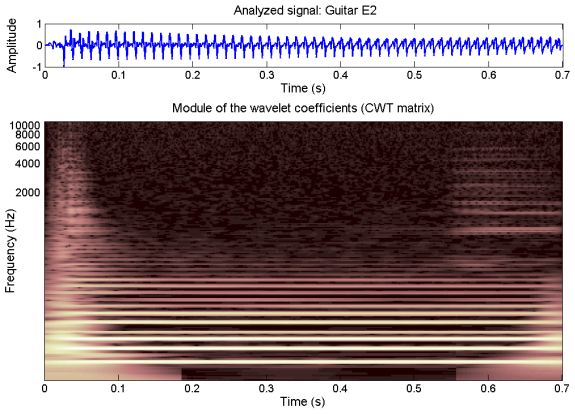

Figure 2: *Up: Waveform of the analyzed signal. In this case, a guitar playing a $E2$ note. Down: The wavelet spectrogram of the signal (module of the CCWT coefficients). The bright zones are the different detected partials.*

the CWAS algorithm, we extend the definition of partial not exclusively to the scalogram peaks, but to their *regions of influence* [13]: for each peak j of the scalogram, the partial P_j contains all the information located between an upper (u_j) and a lower (l_j) frequency limits (which define the region of influence of the peak). This complex function $P_j(t)$ can be written as:

$$P_j(t) = \sum_{m(j)=l_j}^{u_j} W_x(k_{m(j)}, t) \quad \forall j = 1, ..., n \qquad (14)$$

where $W_x(k_{m(i)}, t)$ are the wavelet coefficients related with the i^{th} peak (partial).

The scalogram of the guitar playing an $E2$ note (see Figure 2) can be seen in Figure 3.

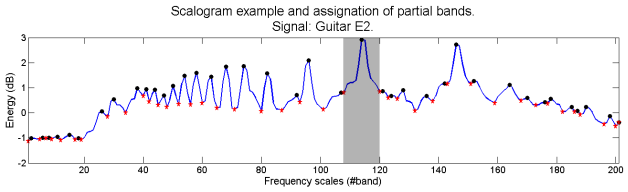

Figure 3: *Scalogram of a guitar playing a $E2$ note. The location of the energy peaks is marked with black circles. The scale* zones of influence *for each partial are marked with red stars. The gray rectangle is the zone of influence of the the second harmonic. The scales or analysis bands are represented in the horizontal axis.*

Studying the complex–valued function $P_j(t)$ in module and phase using Equations (11) and (12), we can obtain the instantaneous amplitude $A_j(t)$ and the instantaneous phase $\Phi_j(t)$ of each detected partial. The instantaneous frequency of the partial can be obtained through Equation (13).

We can obtain the energy of each partial through its amplitude, as:

$$E_j = \sum_{m=1}^{L_j} \|P_j(t_m)\|^2 \qquad (15)$$

where t_m is the m^{th} sample of the temporal duration of the partial j, (whose length is L_j, in samples).

The original signal $x(t)$ can be resynthesized through a simple additive synthesis process, performing the summation of whole set of n detected partials:

$$x(t) = \Re\left(\sum_{j=1}^{n} P_j(t)\right) = \sum_{j=1}^{n} A_j(t)\cos[\Phi_j(t)] \qquad (16)$$

Comparing this expression with Equation (2) we see that the main difference is that in Equation (16), the non-harmonic partials (inharmonic or noisy either) are modeled as the harmonic ones. To obtain the harmonic part of a sound, we use a F_0 estimator presented below.

That is, through the CWAS algorithm and for each detected component, $P_j(t)$, of the input signal, we can obtain:

- Its instantaneous amplitude, $A_j(t)$, through Equation (11).
- Its instantaneous phase $\Phi_j(t)$, through Equation (12).
- Its instantaneous frequency $f_j(t)$, through Equation (13).
- Its energy E_j, through Equation (15).

The CWAS algorithm calculates the wavelet coefficients in frames of 4095 samples. The analysis of these coefficients is performed over accumulated scalograms of 256 samples, tracking the different partials. The multiple F_0 estimation is calculated once by scalogram, hence the accuracy of the proposed technique can be measured along the signal duration.

3. MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION

As was advanced in Section 1, an new algorithm of multipitch analysis has been developed. In this work, we have considered that a musical instrument cannot play more than one note simultaneously (that is, we work mainly with monophonic instruments). If an instrument plays two or more notes simultaneously (polyphony), the developed algorithm will consider that each note comes from a different source. With such an approximation, the present fundamental frequencies F_{0j}, $j = 1, \ldots, N$, become the natural parameter which will be used to calculate the number of sources present in the signal. Figure 4 shows a block diagram of the proposed algorithm.

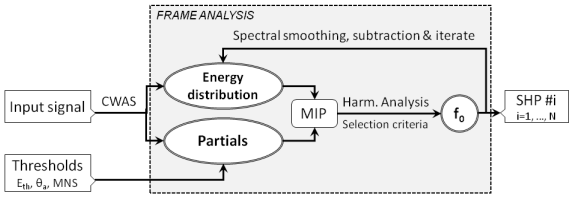

Figure 4: *Block diagram of the fundamental frequencies estimation algorithm.* MIP *is the most important (energetic) partial. The output* SHP#i *is the set of the harmonic partials corresponding to each detected source (*i=1, . . . , N*). See text for details.*

As advanced before, the input signal is analyzed using the CWAS algorithm, which each 256 samples provides the detected partials, tracking them along the signal. As has been detailed, for each partial $P_j(t)$ we know $A_j(t)$, $\Phi_j(t)$, $f_j(t)$ and E_j. We can

easily obtain the average value of the instantaneous frequency for each partial, $\overline{f_j}$:

$$\overline{f_j} = \overline{f_j(t)} = \frac{\sum_{m=1}^{L_j} f_j(t_m)}{L_j} \tag{17}$$

where L_j is, as in Equation (15), the length of the partial (in samples).

Using the mean frequency of each partial and its energy (Equations (17) and (15) respectively), the energy distribution of the signal versus frequency is obtained. This information is the clustered representation of the scalogram of the signal around the discrete set of detected partials.

In order to estimate the candidates to be the (first) fundamental frequency of the signal, only the partials with energy greater than a certain threshold (E_{th}=5%) will be considered in the search of the harmonic sets associated with each source.

In Figure 5 we show the scalogram, between 20Hz and 700Hz, of a frame corresponding to a 6 note polyphony signal (blue line). The energy of all the detected partials is marked with red points while the energetic partials ($E_i > E_{th}$) are marked with black circles. The most energetic partial (MIP) will be used to estimate the first F_0 candidate. In this case, the frequency of this partial is $\overline{f} \approx 726.6$Hz.

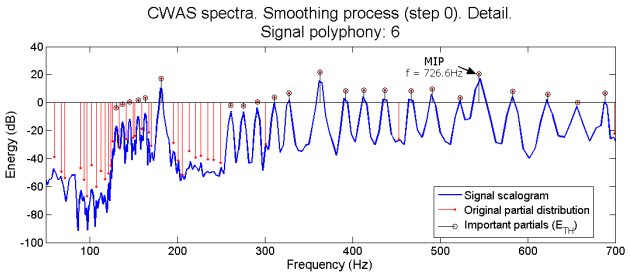

Figure 5: *Scalogram (spectrum) and energy distribution of a 6 note polyphony signal. Detail (up to 700Hz). Blue gross line: original scalogram data. Red dots: energy distribution of the discrete set of detected partials. Black circles: set of important partials (used to estimate the fundamental frequency). The most important partial (MIP) is also marked (see text).*

From the MIP, the harmonic analysis is then computed. Let $\overline{f_j}$ be the average of the instantaneous frequency of the MIP. It is assumed that this partial can be harmonic of a certain fundamental frequency F_{0k}, that is:

$$F_{0k} = \frac{\overline{f_j}}{k}, \qquad \forall k = 1, 2, \ldots, N_A \tag{18}$$

In this work, we have taken $N_A = 10$. In other words, the MIP will be at most the 10^{th} harmonic of its related fundamental frequency. From the candidates to be the fundamental frequency so obtained, the set of harmonic frequencies regarding each one is calculated:

$$f_{k,m} = mF_{0k}, \qquad \forall m = 1, 2, \ldots, N_k \tag{19}$$

where N_k is the higher integer that satisfies $N_k F_{0k} \leq f_s/2$, being f_s the sampling rate.

In the next step, for each $f_{k,m}$, its related partial is searched. A partial of mean frequency $\overline{f_i}$ is the m^{th} harmonic of a certain fundamental frequency F_{0k} if:

$$\left| \frac{\overline{f_i}}{F_{0k}} - \frac{f_{k,m}}{F_{0k}} \right| \leq \theta_a \tag{20}$$

where θ_a is the *harmonicity* threshold. Taking θ_a=0.01, the partials of an inharmonic instrument like the piano [20] are correctly analyzed. This way, the set of harmonic partials (SHP) of each candidate is computed.

The decision of the F_0 associated with the selected MIP is made using three selection criteria:

1. At least one of the first two partials of the SHP must have high energy.
2. The number of energetically important partials within the first 10 harmonics must be as high as possible.
3. The total energy of the SHP must be the highest.

The first two criteria are evaluated first. In doubtful cases (when these criteria offer the same output for different candidates), the third criterium is taken in account. This way we avoid the use of *ad-hoc* weighted functions, which tend to present some kind of signal dependance. The fundamental frequency related to the current MIP is the only winner of these three criteria.

The algorithm stores the set of harmonic partials or spectral pattern, $\mathbf{P}_k = \{P_{1,k}, P_{2,k}, \ldots, P_{n_a,k}\}$ with respective energies $\mathbf{E}_k = \{E_{1,k}, E_{2,k}, \ldots, E_{n_a,k}\}$ (this set includes the obtained fundamental partial, $P_{1,k}$). Over this set, following the proposal of Pertusa and Iñesta [21], we apply a Gaussian spectral smoothing to its energy distribution:

$$\widetilde{E_{i,k}} = G_w \star \widetilde{\mathbf{E}_k} \tag{21}$$

where $G_w = \{0.212, 0.576, 0.212\}$ is a truncated normalized Gaussian window with three components, $\star$ is the convolution product operator and $\widetilde{\mathbf{E}_k}$ is the part of the scalogram centered on each $E_{j,k}$ containing its two closest neighbors. The fundamental partial is erased from the spectrum. The smoothed energy $E'_{i,k}$ for each one of the other harmonic partials is calculated as:

$$E'_{i,k} = \begin{cases} E_{i,k} - \widetilde{E_{i,k}} & \text{if } E_{i,k} - \widetilde{E_{i,k}} > 0 \\ 0 & \text{if } E_{i,k} - \widetilde{E_{i,k}} \leq 0 \end{cases} \tag{22}$$

Substituting these new energy values into its corresponding partials of the original energy distribution, the smoothed energy distribution is calculated, from which a new MIP can be obtained. The process is iterated until the energy of the distribution descends under a threshold (the remaining energy level, E_r = 5%) or the maximum number of sources (MNS in Figure 4) is reached. This information will be used to separate the different sources. As the numerical and acoustical quality of the separation decreases while the number of sources increases, we have limited the number of sources to MNS=10. Repeated F_0 are not taken in account more than once.

Using this technique, it is possible to obtain the fundamental frequencies even in signals with harmonic relations between fundamental frequencies and in the case of suppressed fundamentals. Overlapping fundamentals will not be detected.

3.1. The inharmonic limit

Inharmonicity is a phenomenon occurring mainly in string instruments due to the stiffness of the string and nonrigid terminations. As a result, every partial has a frequency that is higher than the corresponding harmonic value. For example, the inharmonicity equation for a piano can be written [20] as:

$$f_n = nF_0\sqrt{1+\beta n^2} \tag{23}$$

where n is the harmonic number and β is the inharmonicity parameter. In Equation (23), β is assumed constant, although it can be modeled more accurately by a polynomial up to order 7 [22]. It means that the parameter β has different values depending on the partials used to calculate it. Partials situated in the 6-7 octave provide the optimal result. Using two partials of order m (lower) and n (higher), it is:

$$\beta = \frac{\delta - \varepsilon}{\varepsilon n^2 - \delta m^2} \tag{24}$$

where $\delta = (mf_n/nf_m)^2$ and ε is an induced error due to the physical structure of the piano which cannot be evaluated [20]. If partials m and n are correctly selected, $\varepsilon \approx 1$.

With the inharmonic model of Equation (24), it is possible to calculate the inharmonicity parameter β for each detected source, using (when possible) two isolated partials situated in the appropriate octaves. A priori, this technique includes inharmonic instruments (like piano) in the proposed model.

As advanced before, the F_0 of inharmonic instruments like piano can be correctly estimated through an appropriate harmonic threshold θ_a.

4. RESULTS

First of all, we will show the experimental results of this technique evaluated over a 6 note polyphony signal. As the other analyzed signals, this signal have been synthetically generated using the musical instrument samples of the University of Iowa [23][1].

In this signal, the six fundamental frequencies are: $F_{01} = 130$Hz ($C3$), $F_{02} = 137$Hz ($C\#3$), $F_{03} = 145$Hz ($D3$), $F_{04} = 155$Hz ($D\#3$), $F_{05} = 163$Hz ($E3$) and $F_{06} = 181$Hz ($F\#3$). Observe that there are no harmonic relations between these frequencies. The scalogram of a 256 samples frame of this signal is presented in Figure 5 (blue line). The detected partials are also marked, specifically the most important partial of the first step of the analysis (MIP, with a frequency $\overline{f} \approx 726.6$Hz). From this partial, applying the estimation criteria 1 to 3, the fundamental frequency obtained is $f_{01} = 181$Hz$= F_{06}$. The set of harmonic partials (SHP of Figure 4) is smoothed using Equations (22) and (21). The fundamental partial is completely erased from the energy distribution. This information have been presented in Figure 6, where the original scalogram is the grey line, while the original energy distribution of the detected partials is also shown (blue line, each blue dot is the energy of a partial). The red line is the new energy distribution, obtained after the smoothing process. Observe how the fundamental frequency (f_{01}) has been erased, while its harmonics (in the Figure only $2f_{01}$ and $3f_{01}$ are visible) present lower energies.

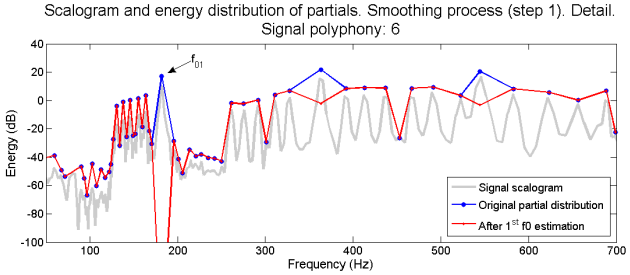

Figure 6: *Blue line: original energy distribution of a signal with 6 notes of polyphony. The energy (dB) of the partials is marked with blue dots. In red (line and dots): the new scalogram, obtained from the first one applying the smoothing process. The fundamental frequency f_{01} is marked with an arrow. The original scalogram is the grey line. Detail (up to 700Hz).*

After this first smoothing, a new MIP is found and the process described in Section 3 is repeated, obtaining a new fundamental and performing a new smoothing. The process is iterative, and ends automatically,as advanced before, when the maximum number of sources (MNS = 10) is reached or when the remaining energy of the signal os lower than a threshold associated with the maximum number of sources (MNS) of the algorithm. In our case, E_r =100/(2 MNS) = 5%. In Figure 7, the scalogram of the original signal, the first smoothed energy distribution (see Figure 6) and the remaining energies after 4 smoothing processes are presented. Observe that most of the partials have been at least partially smoothed, while $f_{02} = 163$Hz$= F_{05}$, $f_{03} = 145$Hz$= F_{03}$ and $f_{04} = 137$Hz$= F_{02}$ have been found.

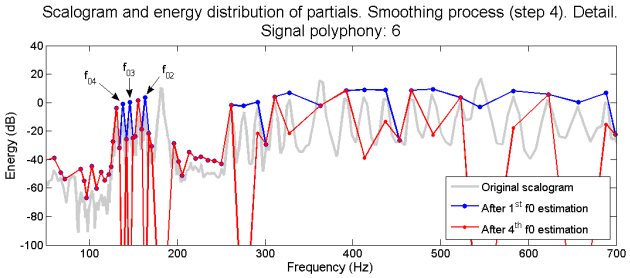

Figure 7: *In blue: energy distribution (dB) of a signal with 6 notes of polyphony after the first smoothing process (see Figure 6). In red (line and dots): the scalogram after 4 smoothing processes. The fundamental frequencies f_{02} to f_{04} are also marked. The original scalogram is the grey line. Detail (up to 700Hz).*

We must remark again that the estimation of F_0 is evaluated each 256 samples of the signal (that is, each 0.012 seconds approximately). The estimated f_{0j} can be stored and plotted in a graphic which shows the temporal evolution of the detected fundamentals. This evolution can be seen in Figure 8 for the 6 notes signal.

In the figure, the trajectories of the 6 detected fundamental frequencies are clearly marked. Observe that the estimation error

[1]Each original archive of the database consists of a certain number of notes played by different musical instruments. Each note is approximately two seconds long and is immediately preceded and followed by ambient silence. The instruments were recorded in an anechoic chamber. Some instruments were recorded with and without vibrato. All samples are in mono, 16 bit, 44.1 kHz, *AIFF* format. We have resampled them at 16 bits, 22.05kHz, *WAV* format. Original excerpts consist of isolated notes. Some of these notes have been synthetically mixed, generating the signals with polyphony 2 to 6.

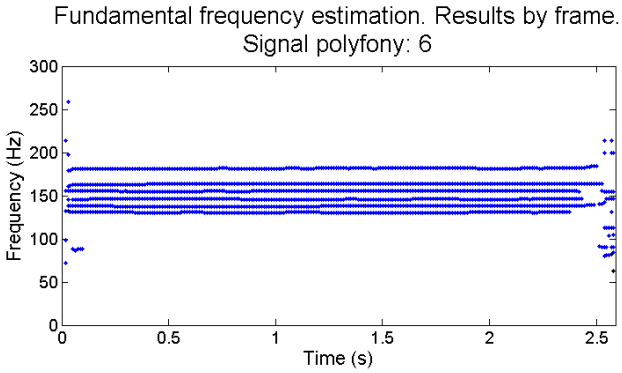

Figure 8: *Evolution of the fundamental frequency estimation along the duration of a signal of 6 notes. The errors tend to concentrate at the beginning and at the end of the signal, where energy is low. The trajectories of the 6 detected (and existing) fundamental frequencies can be easily followed.*

concentrate at the beginning and at the end of the signal. These errors are due to the low energy of the signal during the attack and decay (and/or during the ambient silence before and after the note playing).

As explained before, this algorithm has been tested using a set of signals, synthetically obtained from the musical instrument samples of the University of Iowa [23], which presents 2 to 6 notes of polyphony. The results here shown are preliminar, after the analysis of 13 signals, concretely 3 signals with polyphonies 2 to 4 and 2 signals with polyphonies 5 and 6.

The accuracy of the detection can be numerically evaluated. As the execution of each isolated note can start (and end) in different times, we count the number of successful detections and the number of errors (missing fundamentals or false detections) exclusively where the notes are clearly present (frames 50 to 200, or seconds 0.58 to 2.32 for most of the signals). Proceeding this way, an accuracy percentage measurement can be obtained per signal. These numerical results are shown in Table 1.

In that table, the first column indicates the frequencies involved in the input signal. The numbers of the second column are the theoretical number of detections (that is, the number of frames multiplied by the number of sources). In the third column, the number of experimental detections is shown. In the fourth and fifth columns there appear the number of missed detections and false fundamental detections, respectively. In the last column of the table, the accuracy percentage taking in account the total number of errors. As can be seen, the experimental results show the strength of the technique.

The analyzed signals include harmonic relations between fundamentals, (5^{th} and 12^{th} intervals, minor and major chords). The obtained results are promising.

5. CONCLUSIONS AND FUTURE WORK

In this work, a new estimator of F_0 in polyphonic single-channel signals has been presented. This estimator is based on a modification of the Complex Continuous Wavelet Transform, namely the Complex Wavelet Additive Synthesis, which calculates the wavelet coefficients of the input signal and clusters this information in the scale (frequency) axis, obtaining the partials of the signal. Instantaneous amplitudes, phases and frequencies of the detected partials can be obtained with high accuracy, and this information is used to automatically obtain the set of present fundamental frequencies of a polyphonic signal. The decision of when to stop the iterations is taken using energy criteria or when the maximum number of sources (MNS=10) is reached. This algorithm estimates the different F_0 present in the signal each 256 samples. This way, the F_0 evolution can be easily tracked.

This technique has some limitations. Firs of all, overlapping fundamentals are not detected. Secondly, the F_0 estimator needs a certain energetic presence of the different F_0 to detect it correctly, that is, some detection errors are observed before the attack and after the decay of the present sources. Sometimes false fundamentals can be obtained.

This technique can be improved minimizing the number of errors (missed and false fundamental detections). Two possible improvements are the use of statistical information and a double detection using accumulated information. For example, a fundamental with duration lower than a certain limit could be easily rejected. On the other hand, we can obtain a more energetic scalogram (for example in segments of 4096 samples) and to use it to correct spurious detections.

Once optimized the number of errors, a F_0 tracking algorithm could be easily implemented. Proceeding this way, we could analyze an store the different F_0 present in a larger musical signal, calculating for each note its onset and offset time in order to obtain a first approximation to note detection and segmentation. The set of harmonic partials associated with each estimated F_0 can also be used to separate the isolated sources of the mixture, as presented in [14].

6. REFERENCES

[1] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. ASSP-22, pp. 353–362, 1974.

[2] M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," *Proceedings of the Symposium on Computer Processing Communications.*, pp. 779–797, 1969.

[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[4] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, No. 4, pp. 1917–1930, 2002.

[5] A. Noll, "Cepstrum Pitch Determination," *Journal of the Acoustical Society America*, vol. 41, No. 2, pp. 293–309, 1967.

[6] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, pp. 333–336, 2002.

[7] Y. Medan, E. Yair, and Chazan D., "Super Resolution Pitch Determination of Speech Signals," *IEEE Transactions on Signal Processing*, vol. 39. No.1, pp. 40–48, 1991.

Table 1: *Fundamental frequency estimation results.*

Notes	NoD (Theo.) (#)	NoD (Exp.) (#)	Missed Det. (#)	False Fund. (#)	Accuracy (%)
$C5 + G5$	222	222	0	2	99.1
$C\#4 + E4$	302	300	2	4	97.4
$G\#3 + A\#5$	302	302	0	0	100
$A3 + C4 + E4$	273	273	0	0	100
$C4 + E4 + G4$	423	423	0	3	99.3
$C5 + D5 + E5$	273	273	0	0	100
$C4 + D\#4 + F\#4 + A4$	404	404	0	0	100
$C3 + E3 + G3 + B3$	604	604	0	4	99.3
$B3 + D4 + F4 + G\#4$	364	364	0	0	100
$C4 + D4 + E4 + F\#4 + G\#4$	505	505	0	0	100
$C\#5 + D5 + E5 + F5 + G\#5$	505	505	0	0	100
$C\#4 + D\#4 + F4 + G4 + A4 + B4$	846	838	8	2	97.9
$C3 + C\#3 + D3 + D\#3 + E3 + F\#3$	906	906	0	0	100
TOTAL	**5929**	**5919**	**10**	**15**	**99.6**

[8] P. Cano, "Fundamental Frequency Estimation in the SMS Analysis," in *Proceedings of the Digital Audio Effects Workshop (DAFX'98)*, 1998.

[9] C. Yeh, A. Roebel, and X. Rodet, "Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals," *IEEE Transactions on Audio, Speech, and Languaje Processing*, vol. 18, No. 6, pp. 1116–1126, 2010.

[10] A. Klapuri, "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes," *Proceedings of the 7th International Conference on Music Information Retrieval*, 2006.

[11] A. Klapuri, "Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness," *IEEE Transactions on Speech and Audio Processing*, vol. 11, No. 6, pp. 804–816, 2003.

[12] A. Klapuri, "Multipitch Estimation and Sound Separation by the Spectral Smoothness Principle," *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP'01)*, pp. 3381–3384, 2001.

[13] J. R. Beltrán and J. Ponce de León, "Estimation of the Instantaneous Amplitude and the Instantaneous Frequency of Audio Signals using Complex Wavelets," *Signal Processing*, vol. 90/12, pp. 3093–3109, 2010.

[14] J. Ponce de León, *Análisis y Síntesis de Señales de Audio a traves de la Transformada Wavelet Continua y Compleja: el Algoritmo CWAS*, Ph.D. thesis, University of Zaragoza, Zaragoza, Spain, 2012.

[15] J. R. Beltrán and J. Ponce de León, "Blind Separation of Overlapping Partials in Harmonic Musical Notes Using Amplitude and Phase Reconstruction," *EURASIP Journal on Advances in Signal Processing*, vol. Open Access. doi:10.1186/1687-6180-2012-223, 2012.

[16] I. Daubechies, *Ten Lectures on wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, CBMS-NSF Series Appl. Math., SIAM, 1992.

[17] J. R. Beltrán and J. Ponce de León, "Analysis and Synthesis of Sounds through Complex Bandpass Filterbanks," *Proc. of the 118th Convention of the Audio Engineering Society (AES'05). Preprint 6361*, May. 2005.

[18] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Tran. on Patt. Anal. and Machine Intell.*, vol. 11, no. 7, pp. 674–693, 1989.

[19] B. Boashash, "Estimating and Interpreting the Instantaneous Frequency of a Signal. Part 1: Fundamentals," *Proceedings of the IEEE*, vol. 80, No. 4, no. 4, pp. 520–538, Apr. 1992.

[20] L. I. Ortiz-Berenguer, F. J. Casajús-Quirós, M. Torres-Guijarro, and J. A. Beracoechea, "Piano Transcription Using Pattern Recognition: Aspects on Parameter Extraction," *Proceedings of the 7th Conference on Digital Audio Effects (DAFx'04)*, pp. 212–216, 2004.

[21] A. Pertusa and J. M. Iñesta, "Multiple Fundamental Frequency Estimation Using Gaussian Smoothness," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pp. 105–108, 2008.

[22] L. I. Ortiz-Berenguer, *Identificación Automática de Acordes Musicales*, Ph.D. thesis, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 2002.

[23] L. Fritts and Electronic Music Studios (University of Iowa), "Musical Instrument Samples Database," [Online]. Available: http://theremin.music.uiowa.edu/MIS.html.

[24] S. Amari and J.F. Cardoso, "Blind Source Separation – Semiparametric Statistical Approach," *IEEE Transactions on Signal Processing*, vol. 45, No. 11, pp. 2692–2700, 1997.

[25] A. S. Bregman, *Auditory Scene Analysis: The perceptual organization of sound*, MIT Press, 1990.

[26] G. J. Brown and M. Cooke, "Computational Auditory Scene Analysis," *Computer speech & language, Elsevier*, vol. 8, No. 4, pp. 297–336, 1994.

[27] M. G. Jafari, S. A. Abdallah, M. D. Plumbey, and M. E. Davies, "Sparse Coding for Convolutive Blind Audio Source Separation," *Lecture Notes in Computer Science-Independent Component Analysis and Blind Signal Separation*, vol. 3889, pp. 132–139, 2006.

[28] N. M. Schmidt and Mørup M., "Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation," *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation, ICA'06, Lecture Notes in Computer Science*, vol. 3889, pp. 700–707, 2006.